JOURNAL OF APPLIED MEASUREMENT, 5(1), 15–30

Comparing Traditional and Rasch Analyses of the Mississippi PTSD Scale: Revealing Limitations of Reverse-Scored Items

Kendon J. Conrad
Midwest Center for Health Services and Policy Research, Hines VA Hospital and University of Illinois at Chicago

Benjamin D. Wright
University of Chicago

Patrick McKnight
University of Arizona

Miles McFall
VA Puget Sound Health Care System

Alan Fontana
VA Connecticut Healthcare System
Yale University School of Medicine

Robert Rosenheck
VA Connecticut Healthcare System
Yale University School of Medicine

This study examined whether Rasch analysis could provide more information than true score theory (TST) in determining the usefulness of reverse-scored items in the Mississippi Scale for Posttraumatic Stress Disorder (M-PTSD). Subjects were 803 individuals in inpatient PTSD units at 10 VA sites. TST indicated that the M-PTSD performed well and could be improved slightly by deleting one item. Factor analysis using raw scores indicated that the reverse-scored items formed the second factor and had poor relationships with normally scored items. However, since item-total correlations supported their usefulness, they were kept. The subsequent Rasch analysis indicated that five of the seven worst fitting items were reverse-scored items. We concluded that using reversed items with disturbed patients can cause confusion that reduces reliability. Deleting them improved validity without loss of reliability. The study supports the use of Rasch analysis over TST in health research since it indicated ways to reduce respondent burden while maintaining reliability and improving validity.

Requests for reprints should be sent to Ken Conrad, Midwest Center for Health Services and Policy Research, Hines VA Hospital, Hines, IL 60141, email: kjconrad@uic.edu.

The potential impact of item response theory (IRT) in health services research was previously described in a *Medical Care* supplement (Patrick and Chiang, 2000). However, true score theory (TST) (Nunnally and Bernstein, 1994) remains the dominant psychometric model that is known and used in health services research (McHorney, 1997). Of interest in this study was whether Rasch analysis (Rasch, 1960), could provide more information than TST had provided in answering a psychometric question about the usefulness of reverse-scored items in a measure of posttraumatic stress disorder (PTSD).

Background

The Mississippi Scale for Combat-Related Posttraumatic Stress Disorder (M-PTSD) is a self-report measure that is commonly used for research (Keane, et al., 1997; Litz, et al., 1990; Keane, et al., 1998). It is a 35-item self-report scale derived from Diagnostic and Statistical Manual of Mental Disorders criteria (APA, 1980) for PTSD. Keane, et al. (1997) described a series of three studies designed to explore the psychometric properties of the scale. Of interest here is the first of these that used data from 362 Vietnam veterans seeking help at Vet Centers to confirm the internal consistency of the instrument where Cronbach's alpha was .94. The study also provided a factor analysis where six factors were found, all of which were descriptive of PTSD symptoms. However, eleven of the items were judged not to load on any of these six factors. There was no description of these 11 items.

In another study regarding the factor structure of the M-PTSD, McFall, et al. (1990), in an analysis of 101 combat veterans with PTSD and 101 non-combat veterans with a diagnosis of substance abuse found that 26 of the 35 items contributed to three factors where nine of the items did not load on any factor. Although not noted or analyzed by McFall, et al. (1990), the authors did make clear that these nine were positively worded items that had to be reverse-scored before obtaining a total M-PTSD score. The nine items were left in the scale although their conceptual function in measuring PTSD remained unclear.

Unlike Keane, et al., (1997) and McFall, et al., (1990), King and King (1994) found that the 35 item M-PTSD was unidimensional, i.e., a single-factor solution. However, an IRT analysis by King, et al. (1993) revealed three, positively worded, reverse-scored items that were found to carry very negligible information. This was supportive of McFall, et al., whose analysis indicated the unclear role of the reverse-scored items.

Although the literature on responses to reverse-scored items is sparse, studies by Wright and Masters (1982), Grosse and Wright (1985), Enos (2000), and Bode (2001) have all indicated that the meanings of responses to positively worded items are not the same as the meanings of the negative responses flipped over. In fact, Grosse and Wright (1985) found that response style in the selection of true or false responses can invalidate a total true-false score.

Research Objectives

Three previous studies indicated that the role of a substantial number of M-PTSD items was unclear. In two of these studies, the role of reverse-scored items was unclear, but this finding was not explained, and the items were kept. The principal objective of this study was to analyze a new data set to see if the same problem would be observed and to determine if Rasch analysis could explain it when true score theory (TST) could not. The answer to this question could lead to the desirable outcome of an improved, shorter version (King, et al., 1993; Fontana and Rosenheck, 1994). Brevity accomplished without loss of reliability and validity is an especially worthy objective when at issue is the burden on very troubled patients such as those with PTSD. Improved measurement is a key issue in health services research where imposing burden on ill patients is a common and controversial issue (McHorney, 1997).

Methods

Study Sample

Data were obtained from an observational outcome study of veterans treated in long-term,

short-stay, and brief-treatment PTSD inpatient units at 10 VA sites (Fontana and Rosenheck, 1997). There were 786 to 803 subjects (n of cases varied depending on analysis) who had usable data on the M-PTSD.

All veterans were male. They averaged 45.2 years of age (*SD*=3.2), with 13.0 years of education (*SD*=2.0). In terms of marital status, 38.5% (*N*=309) were currently married and 51.7% (*N*=415) were currently separated or divorced; 9.8% (*N*=79) were either widowed or never married. Ethnically 74.5% were Caucasian (*N*=598), 15.6% were African American (*N*=125), 4.6% were Hispanic (*N*=37), and 5.5% were of other ethnicity (*N*=44).

Scale Description

The M-PTSD Scale contains 35 items rated on a five point response scale, i.e., 1=not at all true; 2=slightly true; 3=somewhat true; 4=very true; and 5=extremely true. Twenty-five items have negative implications, where a higher score indicates greater PTSD, e.g., a score of 5 on "If something happens that reminds me of the military, I become very distressed and upset" indicates greater PTSD. The other 10 items have positive implications, e.g., "I fall asleep easily at night." These 10 positive items indicate an absence of PTSD (Table 1 contains descriptions of items where "+" in the label signifies a positive item). For these positive items, the 12345 rating scale must be reverse-scored to 54321 to remain consistent with the 25 items that indicate presence of PTSD. For example, a 5 on "I fall asleep easily at night" indicates less PTSD so a 5 is rescored as a 1.

Traditional Analysis

The TST analysis began with examination of descriptive statistics and item correlations. The ten items with positive meaning were reverse-scored. Then, this traditional analysis was performed using the SPSS Reliability Program (SPSS, 2001).

After the item statistics and test reliability were examined, a maximum likelihood factor analysis was done to test for the existence of more than one factor. Since the M-PTSD items are summed to yield a total PTSD score, there should be a strong leading factor that is clearly associated with the target construct. A scree plot was used to estimate the number of factors. Then, a varimax rotation was performed to achieve greater clarity on the nature of the factors, i.e., easier to see which items load on which factors. These are straightforward TST methods (McFall, et al., 1990).

Rasch Analysis

The Rasch rating scale model (Wright and Masters, 1982) used for this analysis, estimates the probability that a respondent will choose a particular response category for an item as:

$$\ln \frac{P_{nij}}{P_{ni(j-1)}} = B_n - D_i - F_j$$

where P_{nij} is the probability of respondent n scoring in category j of item i, $P_{ni(j-1)}$ is the probability of respondent n scoring in category j-1 of item i, B_n is the person measure of respondent n, D_i is the difficulty of item i, and F_j is the difficulty of category step j. Rating scale categories are ordered steps on the measurement scale. Completing the j^{th} step can be thought of as choosing the j^{th} alternative over the $(j\text{-}1)^{th}$ in the response to the item (Litz, et al., 1990).

Rasch analysis places persons (B_n) and items (D_i) on the same measurement scale (illustrated in Figure 2) where the unit of measurement is the logit (log odds unit). Person reliability in Rasch is analogous to Cronbach's alpha in TST. It gives an idea of how reliably persons are placed on the scale. Since Rasch places both persons and items on the same scale, reliability can be estimated for items as well as for persons. The Winsteps Computer Program was used for these calculations (Linacre and Wright, 2000). Since reliability estimates are calculated from 0 to 1.00 on scales that are actually infinite in either direction (Linacre, 2002), Rasch analysis provides an alternative statistic, separation. Separation estimates the number of levels from 0 to infinity into which the distribution of persons or items can be reliably distinguished (Smith, E., 2001).

Rather than tailor models to fit the data, Rasch analysis holds that the one parameter model fulfills the requirements of fundamental measurement (Wright, 1997), e.g., linear interval scale, and examines the data, i.e., items and persons, for flaws or problems that are indicated by their failure to fit the model. In this case, the M-PTSD is used as and is therefore assumed to be a unidimensional measure of the construct of posttraumatic stress disorder, and Rasch analysis would be used to test unidimensionality.

Rasch analysis provides fit statistics to test assumptions of fundamental measurement (Wright and Stone, 1979). "Fitting the model" simply means meeting basic assumptions of measurement, e.g., high scorers should endorse or get right almost all of the easy items. Once identified, persons and items that "misfit" can then be examined qualitatively to determine the causes of the problems. Problems may include items with confusing wording or items that assess a construct that is different from the principal one being measured, i.e., multidimensionality. Understanding poor fit can lead to improving or dropping items.

The fit of the data to the model is evaluated by fit statistics that are calculated for both persons and items. The Rasch model provides two indicators of misfit: infit and outfit. These fit statistics have the form of χ^2 statistics divided by their degrees of freedom. The infit is sensitive to unexpected behavior affecting responses to items near the person ability level and the outfit is outlier sensitive. Mean square fit statistics are defined such that the model-specified uniform value of randomness is 1.0 (Wright and Stone, 1979). Person fit indicates the extent to which the person's performance is consistent with the way the items are used by the other respondents. Item fit indicates the extent to which the use of a particular item is consistent with the way the sample respondents have responded to the other items. For this type of analysis, values between .77 and 1.3 are considered acceptable (Smith, R., 2000). In addition to fit statistics, principal component analysis of residuals is used to examine whether a substantial factor exists in the residuals after the primary measurement dimension has been estimated (Linacre, 1998; Smith, E., 2002). The proper functioning of the rating scale is examined using: 1) fit statistics where outfit mean-squares should be less than 2.0, 2) average measures advance monotonically with each category, and 3) step calibrations increase monotonically (Linacre, 1999; 2002; Zhu, 2002; Zhu, Updike, and Lewandowski, 1997). Step calibrations are indicators of the probabilities of categories being observed based on the observed measures of the respondents. Therefore, knowing a respondent's measure should help us to predict what step on the rating scale s/he would choose.

Construct Validation

In Rasch analysis the item hierarchy that is created by the item difficulty estimates provides an indication of construct validity (Smith, E., 2001). The items should form a ladder of low severity symptoms on the bottom to high severity symptoms on the top. In this case, low would involve being startled by noises and having trouble concentrating whereas high would involve despair and suicide.

Another test of construct validity will be the correlation of the original M-PTSD Scale vs. the correlation of a revised M-PTSD Scale with the Clinician Administered PTSD Scale, CAPS (Blake, et al., 1995; Weathers, et al., 2001), the clinical standard for PTSD assessment. The scale with the higher correlation with the CAPS should be regarded as the more valid (Campbell and Fiske, 1959).

Results

Traditional Analysis

Using the SPSS reliability program (SPSS, 2001), Cronbach's alpha was found to be .88 for all 35 items on 786 cases with complete data. In Table 1, which presents the items and their correlations with the total raw score, we found two items, #29 DRUGS at .11 and #2 +GUILT at .18, with item/total correlations less than .2. Only two items, #1 LAKFRNDS and #21 ICRIED, had item/total correlations less than .3. Therefore, DRUGS and +GUILT appeared to be unrelated to the latent construct being measured by the other

items. We examined #29 DRUGS: "There have been times when I used alcohol (or other drugs) to help me sleep or make me forget about things that happened in the military." It is long and confusing since it asks about several issues at the same time, i.e., alcohol, drugs, help sleeping, and making one forget. We concluded that the lack of clarity in the item led to a lack of clarity in its relationship to the construct, so the item was dropped. However, +GUILT, "I do not feel guilt over things that I did in the military" did not have any apparent confusing qualities. It is a reversed item (indicated by the + sign) that is of concern due to previous findings. Based on clear problems with # 29 DRUGS, this item was dropped. The revised 34 item scale had a slightly higher alpha but it still rounded to .88.

Next, we examined the factor structure of the M-PTSD using maximum likelihood extraction. The scree plot (Figure 1) indicated the ex-

Table 1

Item-total Correlations and Cronbach's Alpha Reliability (SPSS). Reverse-coded items begin with "+".

Item #	Label	Items (some abbreviated as indicated by *)	Item Total Corr.	Item Alpha if Item Deleted
1.	LAKFRNDS	Before I entered military, I had more friends.*	.28	.88
2.	+NOGUILT	No guilt over things I did in military.*	.18	.88
3.	PUSHVIOL	If pushed too far, I am likely to become violent.*	.43	.88
4.	BADMLMEM	If reminded of military, I become upset.*	.49	.87
5.	PPLFERME	The people who know me best are afraid of me.	.47	.87
6.	+GETCLOS	I am able to get emotionally close to others.	.34	.88
7.	NGHTMARS	I have nightmares of experiences in military.*	.47	.87
8.	WISHDEAD	Reminded of my deeds, I wish I were dead.*	.52	.87
9.	NOFEEL	It seems as if I have no feelings.	.37	.88
10.	SUICIDE	Lately, I have felt like killing myself.	.40	.88
11.	+SLEPWEL	I sleep well.*	.35	.88
12.	YALIVE	I wonder why I am still alive when others died.	.45	.87
13.	BACKINML	I certain situations, feel I am back in military.*	.46	.88
14.	BADREAMS	Dreams so real, I awaken in cold sweat.*	.50	.87
15.	CANTGOON	I feel like I cannot go on.	.51	.87
16.	DIFEMOTN	Do not laugh or cry at same things as others.*	.45	.87
17.	+ENJOY	I still enjoy doing many things I used to enjoy.*	.42	.88
18.	BADADREM	My daydreams are very real and frightening.	.48	.87
19.	+KEEPJOB	I found it easy to keep job since military.*	.25	.88
20.	TRUBCONC	I have trouble concentrating on tasks.	.44	.88
21.	ICRIED	I have cried for no good reason.	.22	.88
22.	+COMPANY	I enjoy the company of others.	.42	.88
23.	URGES	I am frightened by my urges.	.53	.87
24.	+SLEPEZY	I fall asleep easily at night.	.38	.88
25.	NOISES	Unexpected noises make me jump.	.32	.88
26.	NOUNDERS	No one understands how I feel, not even my family.	.37	.88
27.	+EASYGO	I am an easy-going, even-tempered person.	.39	.88
28.	NOTELL	Things I did I can never tell anyone.*	.41	.88
29.	DRUGS	Used alcohol or drugs to sleep or forget.*	.12	.88
30.	+CROWD	I feel comfortable when I am in a crowd.	.36	.88
31.	LUZCOOL	I lose my cool and explode over minor things.*	.51	.87
32.	SLEPFEAR	I am afraid to go to sleep at night.	.50	.87
33.	AVOIDMEM	I avoid reminders of what happened in military.*	.42	.88
34.	+MEMGOOD	My memory is as good as it ever was.	.32	.88
35.	CANTEXPR	I have a hard time expressing my feelings.*	.43	.88

N of Cases = 786, N of Variables/Items = 35
Reliability Coefficients: Alpha = .8791
Without #29, "drugs" item: Alpha = .8824

istence of two and perhaps three factors: the first with an eigenvalue near 8, a second that was fairly substantial around 2.5 and the third, below 2. These three factors are depicted in Table 2 after a varimax rotation that provided improved interpretability. The first factor was composed of the 25 normally-coded items. The second factor consisted of the 10 reverse-coded items. It was notable that these ten items had negligible loadings on the first factor in the orthogonal rotation. Semantically, they seemed to have little in common except that they were positively worded. In other words, the second factor appeared to be measuring something different from the other 25 items. The third factor dealt with disturbed sleep and dreams but was not a substantial threat to unidimensionality. The key point about the third factor was that, unlike the second factor, the three sleep and dreams items were still related theoretically to the main construct of PTSD.

Although not displayed here because of its large size, the item-to-item correlation matrix helped to enlighten the dimensionality issue. Specifically, the reverse-coded items tended to have low correlations with the other 25 normally scored items, i.e., there were only 7 out of 250 correlations greater than .25, but they had higher correlations among each other ranging from .10 to .58. For example, one would expect +SLEPWEL to be more highly correlated with NGHTMARS (r=.19) and BADREAMS (r=.18) because of theoretical similarity than +SLEPWEL would have with +GETCLOS (r=.37) and +COMPANY (r=.33), but, as seen in the correlations above, this was not the case. Therefore, the 10 reversed items appeared to represent more commonality based on their shared method, i.e., being reverse-scored, than based on the construct they were intended to share with the 25 normally-scored items. This can also be seen in Table 2 where reversed items have high loadings, i.e., high commonality, on the second factor, but very small loadings on the first and third components.

We looked at the item/total correlations again for the reverse-coded items. Since they were reasonably high, i.e., .18 to .42, the usual conclusion would be that they must be contributing to the latent PTSD construct. Obviously, there must be method variance involved, but it did not appear to prevent the items from contributing to the measurement of the latent construct. In contrast, the factor analysis showed that the reversed items did not load on the first factor, the factor that would typically define the principal construct being measured. Instead, they formed their own factor, the second, that appeared to be founded in positive wording or reverse coding. Conceptually this was problematic since the M-PTSD

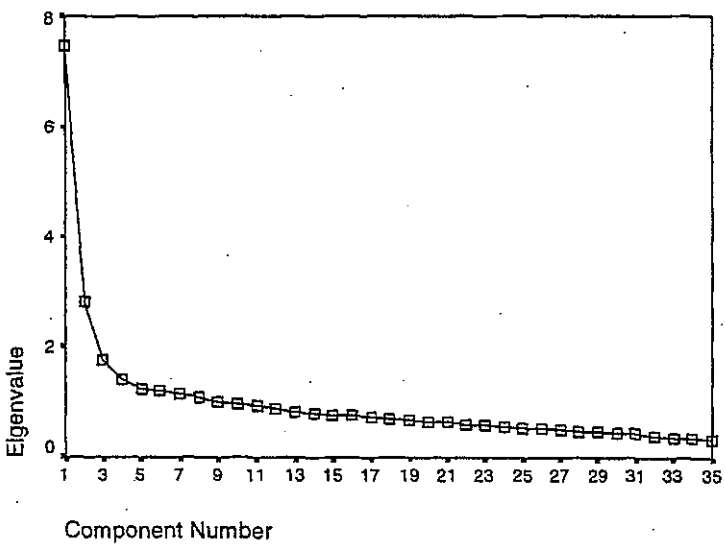


Figure 1. Scree Plot

was designed to have sub-dimensions, but none of these involved reverse-coding.

Another problem with the reversed items was that the items that loaded on the first and third factors had almost nothing in common with the second factor representing reversed items and vice versa. This is indicated by the loadings with negative exponents (E-02) meaning that these loadings are all less than .1. This TST analysis was similar to findings of Keane, et al. (1997) and McFall, et al. (1990), so that we could better understand and appreciate the reasons for their conclusions. We turned to the Rasch analysis to see if it would help to address this seeming contradiction.

Table 2

Rotated Factor Matrix

Variable	Factor 1	Factor 2	Factor 3
LAKFRNDS	.28	.14	.09
+NOGUILT	.07	.22	.04
PUSHVIOL	.39	.19	.19
BADMLMEM	.42	.07	.36
PPLFERME	.54	.11	.15
+GETCLOS	.18	.56	-.07
NGHTMARS	.13	.13	.70
WISHDEAD	.48	.04	.36
NOFEEL	.46	.16	.04
SUICIDE	.46	.02	.19
+SLEPWEL	-.04	.68	.15
YALIVE	.45	.15	.32
BACKINML	.42	-.03	.43
BADREAMS	.16	.12	.75
CANTGOON	.53	.09	.24
DIFEMOTN	.47	.19	.14
+ENJOY	.19	.61	.04
BADADREM	.34	.13	.41
+KEEPJOB	.02	.43	.07
TRUBCONC	.44	.13	.20
ICRIED	.15	.04	.20
+COMPANY	.17	.57	.12
URGES	.55	.10	.26
+SLEPEZY	-.00	.60	.23
NOISES	.19	.05	.35
NOUNDERS	.43	.07	.12
+EASYGO	.37	.32	.00
NOTELL	.29	.08	.36
DRUGS	.15	-.03	.07
+CROWD	.07	.62	.06
LUZCOOL	.54	.17	.16
SLEPFEAR	.27	.16	.56
AVOIDMEM	.28	.14	.35
+MEMGOOD	.15	.47	.00
CANTEXPR	.42	.20	.12

Extraction Method: Maximum Likelihood.
3 factors extracted. 5 iterations required.
Rotation Method: Varimax with Kaiser Normalization.
Rotation converged in 7 iterations.

Rasch Analysis

The analysis of all 35 items on 803 persons produced a person reliability of .86 and separation of 2.46. The Rasch analysis used 17 more cases since it does not require complete data to make valid estimates. Item reliability was .99 and item separation was 10.04. The Rasch person/item map (Figure 2) graphs these results. Reversed items are marked with a plus.

Table 3 provided an examination of the 5-point rating scale. When we looked at the observed average measure for the 5 categories, they were correctly ordered from a low of .24 and progressively higher. However, when we examine the "step calibration," we see a misordering in the step from 2 to 3. Logically, it should be progressively more difficult to go from lower to higher steps on this rating scale. We see, in this case, that it takes less, i.e., less severity of PTSD, to go from step 2 to 3 (-.90) than it did to go from 1 to 2 (-.42).

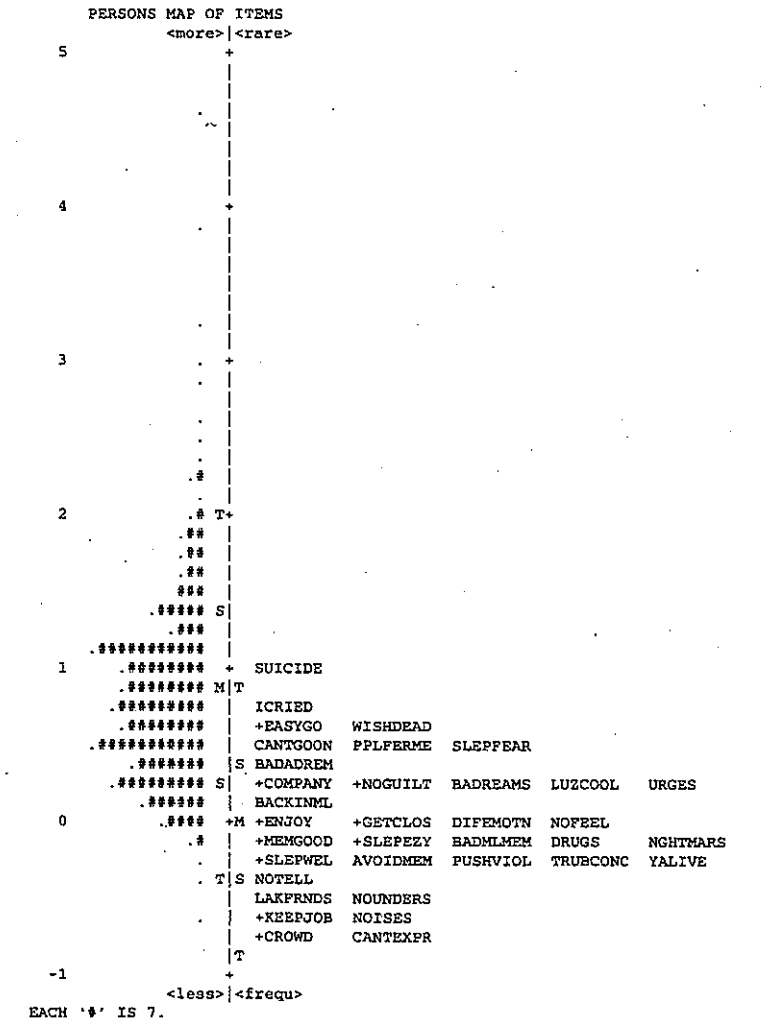
To examine this more closely, Table 4 lists the 6 most misfitting items where both infit and outfit mean squared errors are above 1.3. Item #29 DRUGS, "There have been times when I used alcohol (or other drugs) to help me sleep or make me forget about things that happened in the military" is the worst (infit=1.85, outfit=2.12). This confirms the finding in the TST analysis that this long, complicated item confused respondents.

Four of the six most misfitting items were reverse-scored positive items, i.e., +KEEPJOB, +NOGUILT, +CROWD, +MEMGOOD. This provided further evidence that the reversed items were problematical since they were over-represented in the items with the highest misfit estimates.

To illustrate what poor fit means in this case, Table 5 lists the category measures for the six worst fitting items. Logically, people who choose 1 on an item should, on average, have the lowest overall measure on the M-PTSD. Notice that,

for the worst-fitting, reversed items, that is not true. For #19, +KEEPJOB, category 1 has an average measure of .45 while category 2 has a lower average measure of .43 (asterisks indicate categories that are out of order). This kind of confusion in the response categories exists in all four of these worst-fitting reverse-scored items.

At this point, we may have some evidence to explain why the reversed items tended to misfit (Figure 2). For example, we can examine #30 +CROWD, "I feel comfortable when I am in a crowd" (Table 5). For this item, a score of 1, not at all true, is reversed to a 5 indicating high PTSD. Therefore, subjects with high PTSD should be saying that it is not at all true that they feel comfortable when in a crowd. If the construction of the former sentence seems awkward to you, it must have to the subjects as well since those who answered 4 indicating high PTSD after reversal had an average measure of .52 which was about the same as those who answered 1 who



The numbers on the far left are the Rasch measures in logits.
"M" refers to the person (left side) or item (right side) mean.
"S" is one standard deviation and "T" is two standard deviations.

Figure 2. M-PTSD Person/Item Map of 35 Items

had an average measure of .51. Those who answered 1 should, logically, have had a much lower average measure.

Most unexpected responses. To examine misfit more directly, we listed the 30 most unexpected responses (Table 6). For example, in the first line of Table 6, a response of 4.81 for subject 531 was expected on item #8 WISHDEAD, based on other responses. However, the actual, observed response was a 1 where observed minus expected (1-4.81) leaves a high residual of –3.81. Clearly, the reversed items were over-represented since twenty of the top thirty most unexpected responses were responses to reversed items.

Principal components analysis of residuals. The Rasch principal component analysis of residuals of the 35 items resulted in a strong second factor (or first PC of residuals) composed of the reversed items and "Drugs" (Table 7). This component explained 8.8 of 35 residual variance units or 25%, a substantial amount (Reckase, 1979). This was essentially the same as the SPSS results. Both analyses suggested the existence of a second, "rival" factor represented by the reversed items. In plain terms, this suggested that the rival factor was measuring something different from the principal construct either in terms of content or method. Since all reversed items were included, there is a clear suggestion of a method factor. When we looked at the ten reversed items as a separate measure, the person reliability for this self-contained scale was .66.

The revised scale. Based on the evidence of two misfitting normally scored items, #1 LAKFRNDS and #29 DRUGS and the evidence

Table 3

Summary of Rating Scale Steps for 35 Item Scale

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE MEASURE	
1	1	1175	4	.24	1.37	1.83	NONE	1 not at all true
2	2	1865	7	.17*	.97	1.05	-.42	2 slightly true
3	3	6441	23	.43	.86	.84	-.90	3 somewhat true
4	4	8916	32	.81	.93	.85	.35	4 very true
5	5	9687	34	1.33	.91	.94	.97	5 extremely true
MISSING		21	0	.67				

OBSERVED AVERAGE MEASURE is mean of measures in category. MEASURE refers to the raw score after it has been converted using the Rasch model.
INFIT OR OUTFIT MNSQ (mean squares) below .7 or above 1.3 are regarded as misfitting.

Table 4

Most Misfitting M-PTSD Items

ENTRY NUMBER	RAW SCORE	MEASURE	INFIT MNSQ	OUTFIT MNSQ	SCORE CORR.	ITEMS
29	3235	-.17	1.85	2.12	A .18	DRUGS
19	3475	-.61	2.10	2.04	B .30	+KEEPJOB
2	2923	.28	1.88	1.98	C .27	+NOGUILT
30	3528	-.73	1.66	1.63	D .38	+CROWD
1	3414	-.48	1.61	1.59	E .33	LAKFRNDS
34	3187	-.09	1.38	1.46	F .37	+MEMGOOD

MEASURE refers to the raw score after it has been converted using the Rasch model.
INFIT OR OUTFIT MNSQ (mean squares) below .7 or above 1.3 are regarded as misfitting.

of misfit and questionable construct validity in the ten reversed items, the remaining 23 normally scored items were analyzed. The summary analysis indicated a person reliability of .85 with separation of 2.37 and item reliability of .99 with item separation of 12.21. This was a slight decrease in person reliability and separation and an increase in item separation with item reliability virtually the same. Therefore, the 23 item version, while being shorter, had about the same person

Table 5

M-PTSD Items Category/Option/Distractor Frequencies

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	OUTF MNSQ	ITEM	
29	A	1	52	6	.67	1.9	DRUGS	1 not at all true
		2	44	5	.77	2.0		2 slightly true
		3	125	16	.70*	1.6		3 somewhat true
		4	185	23	.66*	.8		4 very true
		5	396	49	.98	1.1		5 extremely true
		MISSING ***	29	4	1.35			
19	B	5	51	6	.45	1.5	+KEEPJOB	5 not at all true
		4	26	3	.43*	1.2		4 slightly true
		3	78	10	.57	1.2		3 somewhat true
		2	102	13	.61	.6		2 very true
		1	546	68	.97	1.1		1 extremely true
		MISSING ***	28	3				
2	C	5	92	11	.71	1.8	+NOGUILT	5 not at all true
		4	87	11	.61*	1.5		4 slightly true
		3	152	19	.59*	1.0		3 somewhat true
		2	154	19	.78	1.1		2 very true
		1	317	40	1.07	1.2		1 extremely true
		MISSING ***	29	4	1.35			
30	D	5	32	4	.51	1.7	+CROWD	5 not at all true
		4	31	4	.31*	1.1		4 slightly true
		3	42	5	.30*	.7		3 somewhat true
		2	177	22	.52	.5		2 very true
		1	520	65	1.03	1.0		1 extremely true
		MISSING ***	29	4	.83			
1	E	1	45	6	.47	1.5	LAKFRNDS	1 not at all true
		2	17	2	.42*	1.1		2 slightly true
		3	69	9	.47	.9		3 somewhat true
		4	232	29	.68	.9		4 very true
		5	440	55	1.02	1.0		5 extremely true
		MISSING ***	28	3				
34	F	5	33	4	.49	1.5	+MEMGOOD	5 not at all true
		4	56	7	.46*	1.2		4 slightly true
		3	171	21	.60	1.1		3 somewhat true
		2	181	23	.73	1.7		2 very true
		1	361	45	1.09	1.0		1 extremely true
		MISSING ***	29	4	.30			

AVERAGE MEASURE is mean of measures in category. MEASURE refers to the raw score after it has been converted using the Rasch model.
OUTFIT MNSQ (mean squares) below .7 or above 1.3 are regarded as misfitting.

reliability and slightly improved item separation compared to the 35 item version. Using TST, the Cronbach's alpha for the 23 item scale was exactly the same as for the 35 item scale, .88.

The principal component analysis of residuals for the 23 item measure found a strong measurement component, referred to as a yardstick by WINSTEPS, where the yardstick to first factor ratio was 15.3/1. A minor factor still emerged. The items with positive loadings were dominated by sleep and dreams while the items with negative loadings were dominated by lack of feelings and negative urges. However, this factor only explained 2.1 of 23 residual variance units or less than 10%. It is notable that this factor had substantive meaning in contrast to the prior analysis where there was an obvious method factor, i.e., the reversed wording.

Performance of the Rating Scale for 23 Items. In the 23 item M-PTSD, confusion in the rating scale persisted (Table 8). The observed average measures were not out of order. However, the step calibrations from 2-3 were out of order as they had been with 35 items. This indicated confusion in the rating scale for the verbal anchors of 2-slightly true and 3-somewhat true. Semantically, 2 and 3 are similar indicating that combining them would eliminate some confusion

Table 6

30 Most Unexpected Responses

OBSERVED	EXPECTED	RESIDUAL	ITEM	PERSON	ITEM
1	4.81	-3.81	8	531	WISHDEAD
1	4.81	-3.81	30	128	+CROWD
1	4.80	-3.80	1	363	LAKFRNDS
1	4.77	-3.77	33	147	AVOIDMEM
1	4.77	-3.77	11	722	+SLEPWEL
1	4.76	-3.76	29	433	DRUGS
1	4.74	-3.74	30	246	+CROWD
1	4.73	-3.73	19	402	+KEEPJOB
1	4.69	-3.69	19	2	+KEEPJOB
3	4.92	-1.92	5	522	PPLFERME
1	4.66	-3.66	30	824	+CROWD
1	4.66	-3.66	1	247	LAKFRNDS
4	4.98	-.98	34	408	+MEMGOOD
1	4.65	-3.65	19	365	+KEEPJOB
1	4.65	-3.65	30	511	+CROWD
1	4.65	-3.65	35	184	CANTEXPR
1	4.64	-3.64	6	27	+GETCLOS
1	4.64	-3.64	19	330	+KEEPJOB
1	4.64	-3.64	25	227	NOISES
1	4.62	-3.62	2	120	+NOGUILT
2	4.78	-2.78	26	20	NOUNDERS
1	4.62	-3.62	6	159	+GETCLOS
1	4.61	-3.61	34	398	+MEMGOOD
1	4.61	-3.61	21	141	ICRIED
1	4.60	-3.60	19	708	+KEEPJOB
1	4.60	-3.60	30	14	+CROWD
1	4.59	-3.59	30	810	+CROWD
1	4.59	-3.59	30	717	+CROWD
1	4.75	-2.75	30	356	+CROWD
1	4.57	-3.57	19	488	+KEEPJOB

and, perhaps, improve reliability in future administrations of the M-PTSD.

Construct Validity

The construct validity of both the normal and reversed items was supported by the item hierarchy. For normal items, the low severity end was represented by items indicating "a hard time expressing my feelings" and "unexpected noises make me jump." The high severity end was represented by despair, "I feel I cannot go on" and "Lately I have felt like killing myself."

As a traditional indicator of construct validity, we examined the correlation of the 35 item version and the 23 item version with CAPS which is the gold standard clinical assessment of PTSD. Using raw scores (n=782), the 35 item M-PTSD was correlated .43 with the raw score CAPS whereas the 23 item version was correlated .44. Using Rasch measures (n=800), the 35 item ver-

Table 7

Factor 1 From Principal Component Analysis Of Standardized Residual Correlations For Items (Sorted By Loading)

Factor 1 explains 8.8 of 35 residual variance units

FACTOR	LOADING	INFIT MEASURE	OUTFIT MNSQ	ENTRY MNSQ	NUMBER	ITEM
1	.79	.82	.92	.94	A 17	+ENJOY
1	.78	.99	1.65	1.58	B 30	+CROWD
1	.77	.89	1.20	1.20	C 24	+SLEPEZY
1	.77	.90	1.32	1.30	D 11	+SLEPWEL
1	.72	.70	.72	.75	E 22	+COMPANY
1	.71	.77	1.04	1.05	F 6	+GETCLOS
1	.71	1.02	1.80	1.75	G 19	+KEEPJOB
1	.70	.88	1.40	1.43	H 34	+MEMGOOD
1	.44	.61	.99	1.04	I 27	+EASYGO
1	.44	.67	1.51	1.53	J 2	+NOGUILT
1	.01	-.08	1.66	1.74	K 29	DRUGS
1	-.51	.08	.56	.62	a 15	CANTGOON
1	-.51	-.16	.69	.75	b 14	BADREAMS
1	-.50	.13	.86	.95	c 8	WISHDEAD
1	-.49	.03	.70	.74	d 32	SLEPFEAR
1	-.48	-.18	.62	.69	e 23	URGES
1	-.47	-.31	.57	.65	f 13	BACKINML
1	-.47	-.09	.69	.76	g 18	BADADREM
1	-.45	-.49	.63	.71	h 4	BADMLMEM
1	-.44	-.46	.77	.85	i 16	DIFEMOTN
1	-.43	-.49	.75	.81	j 7	NGHTMARS
1	-.43	-.17	.68	.74	k 31	LUZCOOL
1	-.41	.01	.82	.92	l 5	PPLFERME
1	-.40	-.53	.81	.85	m 12	YALIVE
1	-.38	-.78	1.19	1.21	n 28	NOTELL
1	-.38	.64	1.02	1.03	o 10	SUICIDE
1	-.37	-.62	.88	.94	p 33	AVOIDMEM
1	-.36	-.65	.93	1.02	q 20	TRUBCONC
1	-.35	-.67	.77	.84	R 26	NOUNDERS
1	-.33	-.41	.97	1.06	Q 9	NOFEEL
1	-.33	-.67	1.08	1.14	P 3	PUSHVIOL
1	-.30	-.97	.88	.91	O 35	CANTEXPR
1	-.30	-.91	.91	1.00	N 25	NOISES
1	-.24	-.82	1.49	1.65	M 1	LAKFRNDS
1	-.22	.31	1.11	1.25	L 21	ICRIED

sion was correlated .39 with the raw score CAPS while the 23 item version had a correlation of .40. In both cases the shorter version had a slightly higher validity coefficient. The CAPS correlations with Rasch measures are probably lower because of differences in the distributions between Rasch linear M-PTSD measures vs. non-linear, "S-curved" raw CAPS scores. M-PTSD raw scores, however, have the same non-linear, "S-curved" distribution as the CAPS.

Discussion

The Rasch item difficulty measures, fit statistics, principal component analysis of residuals, and construct validation confirm the finding by McFall, et al. (1990) and King, et al. (1993) that the reversed items function differently from the normally scored items. In fact, the Rasch analysis indicated that the reversed items tended to be confusing which caused high misfit values. In the principal component analysis of the 23 well-fitting items, the strong measurement component and the weak first factor in the analysis of residuals indicated that the 23 item M-PTSD was a purer measure of the unitary construct of PTSD. In practical terms, this means that these 23 items worked well together to measure the general construct of PTSD.

Reliability maintained. TST states that reducing the number of items will predictably reduce the reliability of the test, and the Spearman-Brown formula will provide that prediction. If we apply the general S-B formula assuming the original 35- item TST reliability was .88, we would expect the new 23 item measure to have a reliability of .84. When the raw score TST reliability of the 23 item scale was calculated, it was .88 rather than the predicted .84. The Rasch person reliabilities, already noted above, were .86 for 35 items and .85 for 23 items. Therefore, this was clearly a maintenance of reliability when a decrease would be expected by TST.

Improved validity. The maintenance of high reliability was accompanied by an improved validity coefficient when the 23 item version was correlated with the CAPS. Additionally, the examination of fit statistics to eliminate poorly fitting items improved the Rasch item separation in the 23-item version. This indicated a clearer definition of the construct, i.e., improved construct validity, than was present in the 35-item version (Smith, E., 2001).

The person-item map of the 23-item M-PTSD shows that this version was "easy" for the study sample. This can be viewed in two ways. First, it appears that this was indeed a very ill group of patients who readily endorsed the symptoms presented by the M-PTSD. Second, the M-PTSD could be made somewhat more reliable and valid by adding items that indicate higher severity. This same type of issue was observed in an analysis of the CAPS (Betemps, Smith, Baker, Rounds-Kugler, 2003) where the highest severity item was "flashbacks." Betemps, et al. note that achieving better coverage at the high

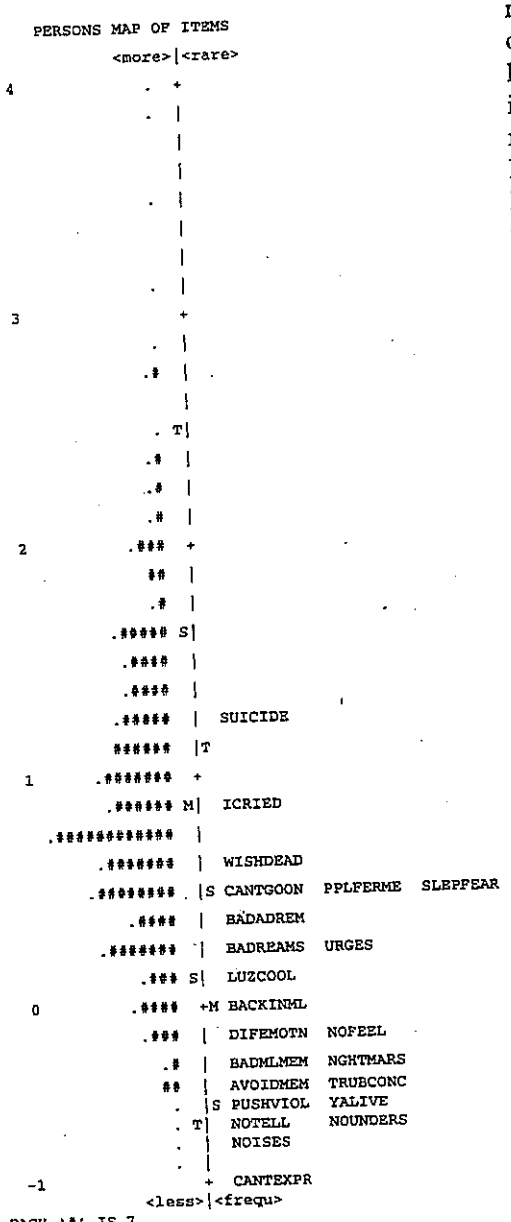
Table 8

Summary of Rating Scale Steps for 23 Items

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	INFIT MNSQ	OUTFIT MNSQ	Step Calib.	
1	1	711	4	-.13	1.29	1.54	NONE	1 not at all true
2	2	1277	7	.00	.98	1.06	-.76	2 slightly true
3	3	4868	26	.44	.92	.92	-1.08	3 somewhat true
4	4	5771	31	.96	.93	.90	.56	4 very true
5	5	5780	31	1.64	.94	.96	1.28	5 extremely true
MISSING		16	0	.83				

Average measure is mean of measures in category.

end of severity is necessary to improve the observation of change in persons with high severity PTSD. The perplexing question then arises: Are there items higher in severity than suicide?



The numbers on the far left are the Rasch measures in logits. "M" refers to the person (left side) or item (right side) mean. "S" is one standard deviation and "T" is two standard deviations.

Figure 3. M-PTSD Person/Item Map of 23 Items

Conclusion

While the TST analysis provided clear evidence of a problem with one item, the Rasch analysis provided clear evidence to support a simplified, shorter, more reliable, more valid, and more user-friendly M-PTSD scale. The analysis of internal consistency and point-biserial correlations provided a good illustration of how these indicators do not provide helpful information regarding the dimensionality of the test (Green, Lissitz, Mulaik, 1977; Hattie, 1985). Rather, the Rasch fit statistics and analysis of residuals were much more informative regarding those items that failed to contribute to the principal measurement dimension. In this study, an examination of the misfitting items revealed improper use of the rating scale that indicated confusion by subjects in answering the reversed items. This supports prior work by R. Smith (1996), who used simulated data and found that Rasch analysis fit statistics with principal component analysis of residuals was generally better than factor analysis except when two factors had about the same number of items.

Additionally, Rasch analysis revealed confusing wording in the rating scale choices themselves that would support a revision from five to four categories. Therefore, in this study, Rasch fit statistics provided useful information beyond that provided by TST. Using TST criteria, the resulting 23-item scale had the same alpha reliability as the 35-item scale and an improved validity coefficient.

Principally, the Rasch analysis provided improved ability to detect confusing wording. These confusions may not be readily apparent to intelligent, healthy researchers and clinicians. However, the confusion of very disturbed patients was reflected in their counter-intuitive responses to difficult constructions. A good final example of this in the M-PTSD was the item, "I do not feel guilt over things that I did in the military" where the response, "not at all true," indicated high PTSD. To answer this correctly, we would expect the most disturbed patients to reverse the most common pattern of their responses, i.e., from

5 to 1, to indicate high PTSD. At the same time, we expect them to use a double negative to indicate a negative quality. In other words, to say, "It is not at all true that I do not feel guilty" indicates that I feel guilt which is a negative feeling. Even healthy people who are feeling well psychologically would have to stop and think about that one.

These findings about reversed items may have greater implications that should be explored more generally in studying clinical as well as non-clinical samples. This study indicated that Rasch analysis is a useful tool to facilitate the achievement of this goal.

References

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Betemps, E. J., Smith, R. M., Baker, D. G., Rounds-Kugler, B. (2003). Measurement precision of the clinician administered PTSD scale (CAPS): A Rasch model analysis. *Journal of Applied Measurement, 4*, 59-69.

Blake, D. D., Weathers, F. W., Nagy, L. M., Kaloupek, D. G., Gusman, F. D., Charney, D. S., and Keane, T. M. (1995). The development of a Clinician-Administered PTSD Scale. *Journal of Traumatic Stress, 8*, 75-90.

Bode, R. K. (2001). Partial credit model and pivot anchoring. *Journal of Applied Measurement, 2*, 78-95.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Enos, M. M. (2000). Just say "No!" *Popular Measurement, 3*, 34-39.

Fontana, A., and Rosenheck, R. (1994). A short form of the Mississippi Scale for measuring change in combat-related PTSD. *Journal of Traumatic Stress, 7*, 407-414.

Fontana, A., and Rosenheck, R. (1997). Effectiveness and cost of the inpatient treatment of posttraumatic stress disorder: Comparison of three models of treatment. *American Journal of Psychiatry, 154*, 758-765.

Green, S. B., Lissitz, R. W., and Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827-833.

Grosse, M. E., and Wright, B. D. (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement, 45*, 1-13.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Keane T. M., Newman, E., and Orsillo, S. M. (1997). Assessment of war-zone related PTSD. In J.P. Wilson and T.M. Keane (Eds.), *Assessing Psychological Trauma and PTSD: A Handbook for Practitioners*. New York: Guilford Press.

Keane, T. M., Caddell, J. M., and Taylor, K. L. (1988). Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: Three studies in reliability and validity. *Journal of Consulting and Clinical Psychology, 56*, 85-90.

King, L. A., and King, D. W. (1994). Latent structure of the Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: Exploratory and higher-order confirmatory factor analyses. *Assessment, 1*, 275-291.

King, D. W., King, L. A., Fairbank, J. A., Schlenger, W. E., and Surface, C. R. (1993). Enhancing the precision of the Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: An application of item response theory. *Psychological Assessment, 3*, 457-471.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*, 266-283.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103-122.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85-106.

Linacre, J. M., and Wright, B. D. (2000). *Winsteps*. Chicago: MESA Press.

Litz, B.T., Penk, W.E., Gerardi, R., and Keane, T.M. (1990). Behavioral assessment of PTSD. In P. Saigh (Ed.), *Post-traumatic Stress Disorder: A Behavioral Approach to Assessment and Treatment*. Boston: Allyn and Bacon.

McFall, M. E., Smith, D. E., Mackay, P. W., and Tarver, D. J. (1990). Reliability and validity of Mississippi Scale for Combat-related Posttraumatic Stress Disorder. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 2, 114-121.

McHorney, C. (1997). Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. *Annals Internal Medicine, 127*, 743-750.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

Patrick, D. L., and Chiang, Y. (2000). Health Outcomes Methodology: Symposium Proceedings. *Medical Care, 38*(suppl II), II-73-II-82.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedogogiske Institut. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*, 281-311.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205-231.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling, 3*, 25-40.

Smith, R. M. (2000). *Common oversights in Rasch studies*. MESA note 9. Retrieved from http://www.rasch.org/rn9.htm.

SPSS, Inc. (2001). *SPSS for Windows*. Chicago: Author.

Weathers, F. W., Keane, T. M., and Davidson, J. R. T. (2001). Clinician-Administered PTSD Scale: A review of the first ten years of research. *Depression and Anxiety, 13*, 132-156.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, Winter*, 33-52.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., and Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.